Comparative Sequence Analysis of Human and Mouse Promoter Regions

Hirokazu Chiba
hchiba@ims.u-tokyo.ac.jp

Riu Yamashita
ryamasi@ims.u-tokyo.ac.jp

Kengo Kinoshita
kino@ims.u-tokyo.ac.jp

Kenta Nakai
knakai@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: promoter region, comparative analysis, alignment

1 Introduction

The DNA sequence just adjacent to the transcriptional start sites (TSSs), what we call the promoter, plays an important role in transcription regulation. Although many attempts have been made to detect functional signals in promoter regions aiming to understand the mechanisms of transcription regulation, those attempts remain challenges because the grammars or semantics of promoter sequences are not fully unveiled. The completion of genome sequences has made the large scale genomic comparison an effective tool for mining the signals from genome sequences. In addition, experimentally validated TSS data has been accumulating in these years [1], which made possible the comparisons of the promoter sequences [2]. The TSS data are especially abundant for human and mouse. In this study, we carried out the most comprehensive comparison to date for promoter regions of human and mouse, and the relationship between gene function and promoter conservation were examined.

2 Method and Results

TSS data of human and mouse were obtained from DBTSS [1]. We extracted 1200 bp of promoter sequences covering from 1000 bp upstream to 200 bp downstream of the TSSs. An ortholog table of human and mouse was also obtained from DBTSS. Using these data, 9400 pairs of promoter sequences were associated as orthologous promoters. They were aligned by a local alignment program LALIGN (a part of fasta2 program package [4]) with the default settings. Promoters of non-orthologous genes were paired between human and mouse and aligned in the same way to be used as negative controls. The results were shown in Figure 1A. The distribution of the alignment scores of orthologous promoters were depicted in solid line, which has two peaks; major peak between 1000 and 2000, and minor peak around 100. Comparing this with the distribution of negative controls, it seems that the minor peak of the orthologous pairs corresponds to the peak of negative controls. We assumed that those pairs showing scores less than 300 are improperly paired sequences. The resulting alignment scores vary in a wide range from 300 to 5203.

We subsequently examined the relationship between gene function and promoter conservation. Annotation of genes was made by associating human genes with GO slim [5] term. The alignment scores of genes with a certain GO slim term were compared with that without the term. In comparisons, we did not use the promoter pairs with an alignment score less than 300. An example of these distributions was shown in Figure 1B, which is in the case of “developmental process”. The difference of the distributions was tested by Wilcoxon rank sum test. Part of the results were shown in Table 1. In the aspect of molecular function, the highest score was observed for transcription regulator, and lowest score for enzyme, and in the aspect of biological process, the highest for developmental process, lowest score for response to stress. The detailed results will be shown in the poster presentation.
Table 1: GO slim terms with significant degree of conservation in promoter regions

<table>
<thead>
<tr>
<th>GO category</th>
<th>GO slim term</th>
<th>Number of genes</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>High score</td>
<td>biological process</td>
<td>910</td>
<td>&lt;10^{-15}</td>
</tr>
<tr>
<td></td>
<td>molecular function</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>developmental process</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>transcription regulator</td>
<td>561</td>
<td>&lt;10^{-15}</td>
</tr>
<tr>
<td>Low score</td>
<td>biological process</td>
<td>477</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>molecular function</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stress response</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>enzyme</td>
<td>2711</td>
<td>&lt;10^{-9}</td>
</tr>
</tbody>
</table>

3 Discussion

The most comprehensive comparison was carried out for promoter regions of human and mouse. The conservation rate of promoters vary in a wide range. And the relationship between gene function and promoter conservation were examined. For the genes involved in “developmental process” or “transcription regulator”, promoter sequences were highly conserved. This result is consistent with the previous reports [2, 3]. And we found out that the promoter sequences were less conserved for the genes associated with “stress response” or “enzyme”. These results suggest that the regulations of genes involved in development or transcription regulation are finely tuned, while the regulations of enzymes or genes involved in stress response are relatively simple.

References


